An introduction to text mining with AntLab and Voyant Tools

Article (Published Version)

An introduction to text mining with AntLab and Voyant Tools

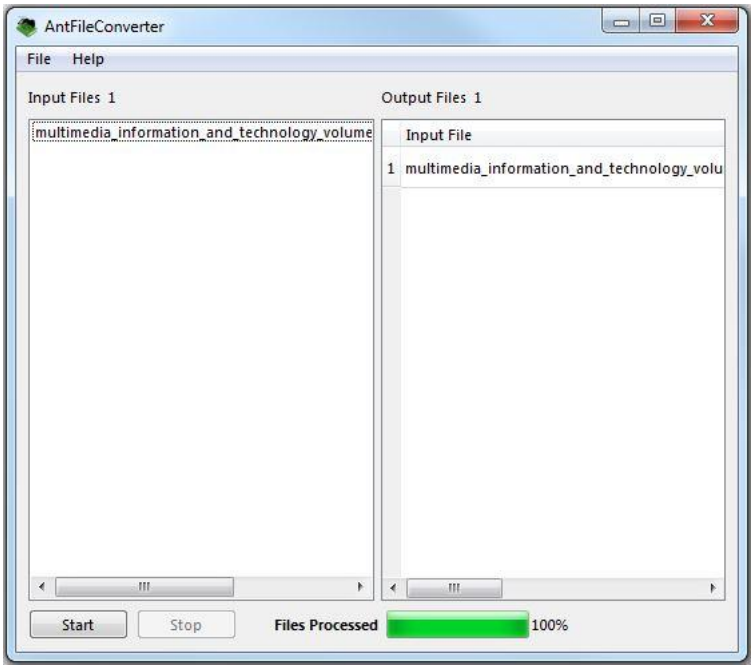
Increasingly you may hear researchers, librarians and other information professionals talk about "text mining". Although this is a process aligned with information retrieval, it is not always clear how we can support and engage with these related activities. The following post brings together a number of resources that show the value and benefits of text mining, and introduces two free tools to help you start exploring this growing area of work.

The introduction to the PLOS Text Mining Collection, a useful selection of open access research and essays relating to text mining, explains that:

> "The maturing field of Text Mining aims to solve problems concerning the retrieval, extraction and analysis of unstructured information in digital text, and revolutionize how scientists access and interpret data that might otherwise remain buried in the literature".

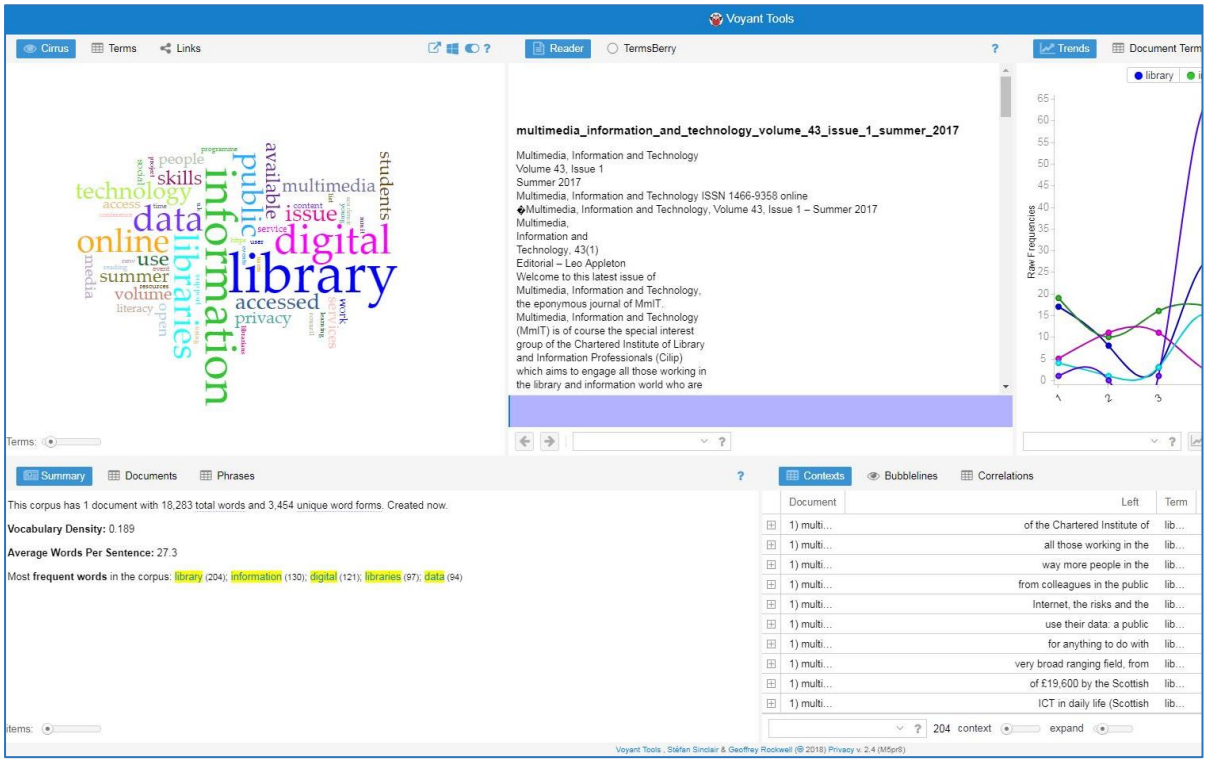
An example of this is Yale University's Robots Reading Vogue project where a huge volume of text and data (over 6 TB) has been analysed to show, amongst other things, how the use of particular words has risen and fallen over the publication's history (the *n*-gram Search). At the University of Sussex there are numerous projects coming from the Text Analysis Group and the Sussex Humanities Lab exploring large corpora (collections of written text by particular authors or about particular subjects) through text mining. We have even started to run workshops in the Library introducing tools to help students who are interested in this area of research. I would like to share two of these resources here: AntLab and Voyant Tools (you can find even more in the TAPoR collection).

AntLab contains a number of freely available tools (although donations and patronage are welcome) built by Dr Laurence Anthony, which can be found on the Software section of his website. For the purpose of this post, I would like to highlight AntFileConverter, a tool for converting PDF and Word files into plain text for analysis - something that can also be helpful for improving accessibility. To use AntFileConverter download and open the appropriate software version for your computer, drag the file you wish to convert into the 'Input Files' box, and click 'Start'. For this demonstration I have used the PDF of the first Open Access volume of the MmIT Journal:



As explained in the user support, “the converted files will be saved in the same directory as the original files with the same name but with the ".txt" extension added”. This *.txt* file can then be used with other AntLab software, although here will be analysed with Voyant Tools, a free “web-based reading and analysis environment for digital texts”. To do this, upload the *.txt* file created with AntFileConverter into the Voyant Tools box:

Click on ‘Reveal’ to run the analysis and view the results:



The default tools include *Cirrus*, *Reader*, *Trends*, *Summary* and *Contexts*, which you can learn more about in the Getting Started Guide. There are also a number of additional tools, including the TermsBerry. To use this particular tool, click on *TermsBerry* next to *Reader* above the second panel:

The TermsBerry shows how often particular terms occur and how frequently they appear next to other terms. The TermsBerry I have shared above shows that in Volume 43 of the MmIT Journal, the words 'library' and 'information' are two of the most common (they are in larger bubbles). If you hover over one of the terms, for example 'digital', you will see that this word appears 121 times in the text, most commonly co-occurring with 'literacy' (29 times), followed by 'skills', 'media' and 'information'; topics that should interest MmIT readers!

To enable this mining and sharing, reforms to Copyright legislation mean that copies of a work can be made for the purposes of text and data analysis (providing you have lawful access to the original work, which in this case is open access). Additionally, as explained in the 'Sharing outputs' section of this Jisc guide, the results of the analysis can usually be shared with anyone (although there are exceptions to this when the analysis goes beyond counts and 'facts' about the work, and includes large amounts of the original copyright material). So armed with a few tools, and copyright law on our side, it's time to make text mining yours.